Dataset for Temporal Analysis of English-French Cognates

Esteban Frossard[♠] Mickaël Coustaty[♠] Antoine Doucet[♠] Adam Jatowt[♢] Simon Hengchen[♣]

[♠]University of La Rochelle, L3i Laboratory, {firstname.lastname}@univ-lr.fr
[♢]Kyoto University, adam@dl.kuis.kyoto-u.ac.jp
[♣]University of Helsinki, simon.hengchen@helsinki.fi

Abstract

Languages change over time and, thanks to the abundance of digital corpora, their evolutionary analysis using computational techniques has recently gained much research attention. In this paper, we focus on creating a dataset to support investigating the similarity in evolution between different languages. We look in particular into the similarities and differences between the use of corresponding words across time in English and French, two languages from different linguistic families yet with shared syntax and close contact. For this we select a set of cognates in both languages and study their frequency changes and correlations over time. We propose a new dataset for computational approaches of synchronized diachronic investigation of language pairs, and subsequently show novel findings stemming from the cognate-focused diachronic comparison of the two chosen languages. To the best of our knowledge, the present study is the first in the literature to use computational approaches and large data to make a cross-language diachronic analysis.

Keywords: Crosslingual semantic change, cognates, temporal analysis, semantic analysis

1. Introduction

Languages, our main tools of communication, evolve constantly: words obtain new and lose old meanings over time, they become popular or fade into obscurity. Because of its importance, language is studied by academics and public alike, as shown by the large number of publications and websites devoted to language evolution, etymology and semantic changes (Cresswell, 2010; Ayto, 2011; Lewis, 2013). Most of these focus on individual words only or are done on a small scale, mainly because the analysis requires manual work to locate occurrences of features in old texts, and then to compare manually their contexts or other characteristics.

In the recent years, large amounts of digitized old books and texts were made available, such as Google's Books initiative (Michel et al., 2010) with 5% of books ever published. Computational approaches have also been conducted to analyze them (Gulordava and Baroni, 2011), proposing novel approaches for understanding lexical semantic change – for an overview, we refer to the survey by (Tahmasebi et al., 2018). However, to the best of our knowledge, no cross-language temporal analysis has been proposed in the literature using computational approaches and large data. In addition, most prior studies focused only on English, whereas comparing two or more languages can shed light on how they actually co-evolved over time.

To study multiple languages over time, we assume the most intuitive approach: we focus on their similar connecting aspects. We use in particular words in both languages that have the same origins and similar meaning, also known as cognate words. We propose to study the temporal characteristics of cognate words as an approach to cross-language diachronic analysis. These cognates, loanwords included (i.e., words that come directly from the other languages) are an important subset of the lexicon and have been frequently studied. Most prior works focused on synchronous analysis of cognates (see for example (Uban et al., 2019)), while we look at their temporal aspects and correlations.

We have used the largest multilingual corpora available on a relatively long time, allowing thanks to its size to set a yearly granularity of analysis. In particular, we used Google Books Ngrams[1] in English and French to conduct the analysis. Despite its inherent problems (Pechenick et al., 2015), it is one of the few corpora of this size available in both French and English. We also prepared a list of English-French cognates based on existing lists and few selection criteria described below.

Cognates are, in linguistics, words that share a common etymological origin (Crystal, 2011), of which loanwords (words borrowed from other languages, e.g. English `communiqué` is borrowed from the French) are particular cases. Both are of great interest in multi-language analysis thanks to the ease of understanding and the identification of links between languages.

Numerous works have focused on either cognates or loanwords. On the one hand there are works for cognate detection harnessing computational methods that propose the first step in a (semi-) automatic analysis of cognates using the vast amount of digitally available data, when manual annotation requires a lot of man-hours (Jäger et al., 2017; List et al., 2018). On the other hand there are semantic analyses of cognates, that manually investigate cognates to look for links between two different languages (List et al., 2018; Aske, 2015). Some recent works cope with the limitations of these two categories by mixing the use of automatic detection of cognates with the semantic analysis (List et al., 2018; Rabinovich et al., 2018).

Nevertheless, to the best of our knowledge, there has been no automatic study of the frequency correlations and patterns of cognates over time across different languages, especially one that uses large datasets. In this paper, we propose a statistical change-oriented analysis of cognates, and focus on English and French.

[1] `https://books.google.com/ngrams`, accessed on November 15, 2019

2. Datasets

We started the study of English-French cognate by constructing a large cognate dataset that fits our criteria (see Section 2.1.). First, we created a list of cognates applicable for our study, basing our selection on available English and French lists of cognates (Bergsma and Kondrak, 2007), removing those that did not fit our criteria and adding some other. Each word's "cognateness" was confirmed by investigating its etymology with the Oxford English Dictionary, the on-line etymology dictionary[2] and the French National Center for Textual and Lexical Resources (FR: *Centre National de Ressources Textuelles et Lexicales*).

We used the 1-gram from the Google Books n-grams, for English and French (Michel et al., 2010) as an underlying dataset. It contains around half a trillion English words and one hundred billion French ones coming from books of varying literature genres. We note that although the dataset is not balanced in terms of document types its strong advantage lies in the very large size in comparison to other similar datasets, both in number of words and periods covered (from the 1500s to the late 2000s).

Finally, we would like to mention that we first focus on the differences in use frequency of words over time, hence we chose Google Books 1-grams. However, the underlying dataset can be easily extended by using larger n-grams such as 5-grams.

2.1. Criteria for Selecting Words

We chose English-French word pairs for constructing the cognates dataset and we based the selection on four criteria as follow. (1) We restricted the time scope to the years from 1800 to 2008, where most of the data is. (2) We chose words that were cognate pairs based on their etymology to make sure they were actual cognates. (3) We discarded verbs as their many inflections in French introduce noise, mostly as shared surface forms with other lexical items. (4) Finally, we chose words that appeared above a minimal frequency threshold (one in two million, or from 35 to 10,000 appearances in a single year, depending on the number of words available for that year) in both English and French to allow a proper analysis and to minimize the chance of an erroneous detection.

Once all words were selected, every inflection of each word was found using dedicated dictionaries. The frequency of all forms of a word were summed for each year to compute the total frequency of the word for that year. We then obtained for each word a time series from 1800 to 2008 representing its frequency. Finally, for each word, the time series, year of the first appearance, the maximum frequency and its year are all stored in a text file.

2.2. Cognates Dataset

Based on the data and the criteria presented above, we built, and release, a cognate dataset with 492 word pairs composed of nouns, adjectives and adverbs[3]. Each pair has between one and four forms in English, and up to ten in French. In English, most words have only one form for adjectives and adverbs, while most nouns have two forms (singular and plural). In French, with masculine and feminine, singular and plural forms, most nouns and adjectives can be found in four different surface forms.

The dataset includes 353 (71%) French loanwords (French words used in English) and 15 (3%) English loanwords[4]. These numbers include words taken from Old French and Old English. Note that the words are eclectic, both in meaning, as we aimed not to bias the dataset to any topic, and in frequency, as shown in Figure 1 where we plot median frequency as well as quartiles.

In the end, the dataset contains, for each cognate, both in English and in French, its frequency all inflexions combined in each year from 1800 to 2008 (0 in years before they appear or they are not part of the dataset).

3. Temporal Analysis of Cognates

We present below the preliminary results of the frequency analysis using the constructed cognate dataset.

3.1. Correlation of Cognates

First, we wanted to examine if the level of use of words in each of the languages changed in their own way or, rather, if the cognates shared similar patterns of changes in the intensity of their use over time. We then started by computing the frequency correlation for each pair of cognates. We used Pearson correlation coefficient(Pearson, 1895) on the time series representing cognate use in the concerned period. The frequency of a term in a given year is computed by dividing the number of occurrences of the term (the sum of the number of occurrences of each of its forms) by the total number of summed appearances of all words in this year.

As shown on Figure 2, there was a strong positive correlation for most pairs, with more than half (57%, 281) having a correlation value above 0.5, and over 13% (65) above 0.9. However, the high positive correlation is not true for every pair, as correlations go from -0.87 for the pair `employee - employé` to 0.99 for the pair `traditionally - traditionnellement`. Nevertheless, the number of pairs with a negative correlation, or close to zero, is rather small, as shown on Figure 2. This suggests that *cognates do not only share a past (etymological roots), but they also share similar usage patterns over time*.

Most of the cognate pairs had correlated changes of frequency over time. On the left of Figure 2, negatively correlated words are quite rare (6%, 31 words below -0.3). This suggests that cases when cognate words have tendencies to change the frequency of their use in an opposite way are quite rare.

If we restrict the analysis to the French loanwords (see the red plot in Figure 2), the positive correlation is similar, 201 loanwords (57%) having a correlation above 0.5 with their counterpart and 46 (13%) having the correlation value above 0.9.

[2] Available online at `https://www.etymonline.com/`

[3] Available online at `https://zenodo.org/record/3688087`.

[4] Due to the small number of English loanwords, we will focus only on French loanwords in our analysis.

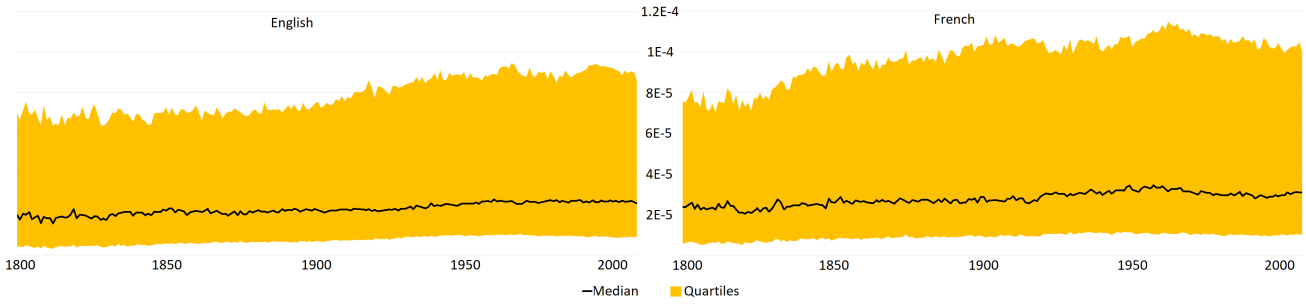

Figure 1: Distribution of the frequencies of cognates pairs, expressed through the quartiles and median.

3.2. Level of Word Use

The correlation of fluctuations in word frequencies over time as studied above still does not tell us whether words were actually used at the similar intensity levels in the same years. One word in a cognate pair could be used very frequently, while its counterpart could be barely used even though their relative frequency changes over time may be correlated.

To compare whether the frequency of a word is similar to its cognate counterpart, we first looked at the ratio between their maximal and mean frequencies. Then, for a cognate pair (w_E, w_F), with $f_E(w, y)$ and $f_F(w, y)$ denoting the frequency (respectively, in English and French) of the word w in year y, we computed the following formula:

$$\frac{max(max_{y\in[1800;2008]} f_E(w_E, y), max_{y\in[1800;2008]} f_F(w_F, y))}{min(max_{y\in[1800;2008]} f_E(w_E, y), max_{y\in[1800;2008]} f_F(w_F, y))}$$

This equation gives a real number of one or greater and is based on the comparison of the maximum frequencies of cognates. The closer to one, the greater the similarity between the maximum frequencies of the two cognates, with the limit at one where both the values (maximum frequency in English and maximum frequency in French between 1800 and 2008) being equal. When the resulting value is higher, the two words in a given cognate pair have a less similar use.

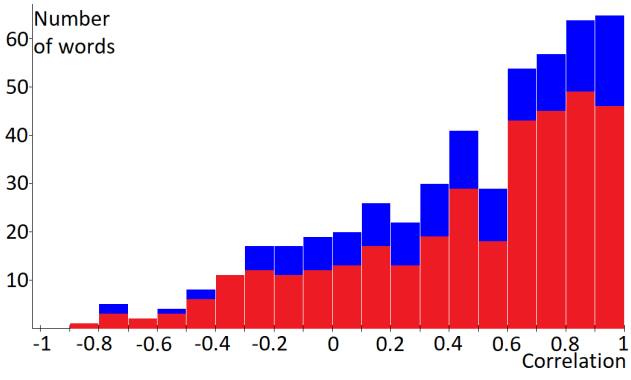

Figure 2: Correlation of English-French cognate pairs (blue) and French loanwords (red), from the first appearance of a word (English or French, depending on the earliest one) to 2008, as including earlier years would artificially increase correlation.

The cognate words not only tend to be correlated in terms of their changes over time, but they also have (for most of them) a similar level of use in their languages. The maximum usage of the most used word in each cognate pair is, for more than half of the words, at most 1.63 times more than its counterpart in the other language.

Moreover, the more we focus on the correlated words, the smaller this median line is (1.53 for correlation above 0.5; 1.49 for correlation above 0.7; 1.48 for correlation above 0.9). If we analyze only the loanwords, the results are similar.

To see if this ratio changes according to the frequency in one or both languages, and if one language has the cognates consistently more used (especially interesting are outliers), their respective mean frequencies seem to follow a linear distribution (see Figure 3). However, there are also cases of high frequency of use of a cognate in one language with low frequency in the other language (even several thousand times more in one language).

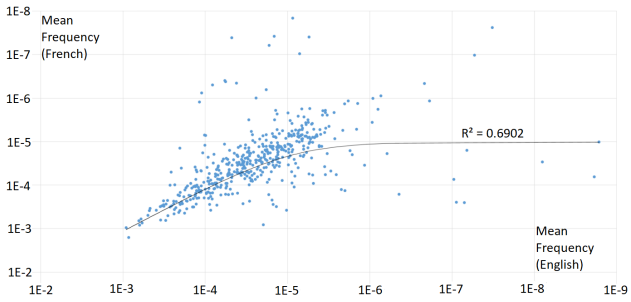

Figure 3: Distribution of the mean frequency in French according to the mean frequency in English (log-log plot). The linear regression $y = 1.1457x + 10^{-5}$ (black) shows the global relation between mean frequencies.

These extremes tend to be as likely to result from higher use in English as in French. As the correlation analysis indicated that the level of use of cognates evolved according to the same pattern across time, the frequency ratio indicates the *cognates have a similar level of use in both languages across time*.

3.3. Language Specificities

As the results show that cognate words are often used similarly at the same time in both the languages, one could be

tempted to say that a cognate, independently of language, performs in general a similar role in both languages and is used in very similar ways over time.

There are several potential reasons that could be proposed behind the differences in use frequencies and their temporal variations over time in both languages. To a certain degree, these could be explained by the subtle differences in the meaning of the cognates in both the languages, which would be used for slightly different purposes or in differing situations. Another driving force behind the observed differences in cognate use could be the existence of a synonym or multiple synonyms in only one of the two languages, which could "drain" the usage of one of the two words of the cognate pair: as per (Saussure, 1916), there is no bijective relationship between words in different languages.

Another explanation could be the occurrence of an additional acquired sense behind a cognate in one language increasing the use of this word with relation to its use in the other language. For example `azote` is barely used in English, in favor of `nitrogen`, while it is the opposite in French (`nitrogène` exists, yet `azote` is more commonly used).

3.4. Impact of External Factors

French and English are not only affected by each other, but by a multitude of external factors which can explain at least some of the correlations between cognates pairs, like the common history of corresponding countries. Analyzing history – i.e., the context around language use – can lead to an understanding of the impact of important events on some words, the most explicit example in our dataset being `bombardment - bombardement`, shown in Figure 4, a word which was obviously used more frequently in times of war, or, rather in the case of our corpus, when war-related books were popular. However, such effects are often difficult to determine, especially when the causes are less known.

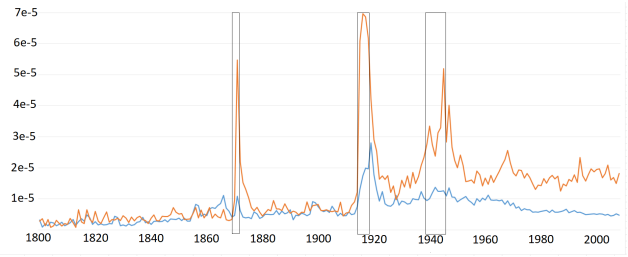

Figure 4: Frequency of `Bombardment` (English, in blue) and `Bombardement` (French, in orange) from 1800 to 2008. Three spikes can be observed (denoted by black rectangles), which correspond to the Franco-Prussian war (1870-1871), World War I (1914-1918) and World War II (1939-1945), showing the effect of the events on the languages.

4. Limitations

The dataset is not exempt from limitations, from its rather small size, as we focused on most-known cognates for the first analysis, to potential bias coming from the choice of words, even if we did our best to limit it, or from the corpus choice. We also provide the results of preliminary frequency-focused analysis of the cognates based on the created dataset. The analysis itself has some limitations: as it only covers two well-known languages, English and French, and only by not taking into accounts synonyms that made some cognates out of use in one of the two languages.

5. Conclusions & Future Work

In this paper, we describe a dataset of English and French cognates constructed to study their evolution from 1800 to 2008.

Diachronic language analysis and in particular studies of word origins have recently attracted considerable attention. In this paper we also emphasized the idea of studying temporal variability of a language by its synchronized comparison with another language where the synchronization is based on using cognates (serving as a comparative "bridges") aligned over time. By this, we add a second dimension or an additional investigation axis to the usual diachronic analysis approaches.

In the future, we plan to extend the current study to embrace larger number of cognates and to conduct a semantic analysis of the cognate variation across time and languages. We will also study other language pairs including ones that had less interaction and exchange in the past.

6. Acknowledgments

This work has been supported by the European Union Horizon 2020 research and innovation programme under grants 825153 (Embeddia) and 770299 (NewsEye).

7. Bibliographical References

Aske, J. (2015). Spanish-English cognates: An introduction to Spanish linguistics. Open Access eBook (Open Textbook). CC BY-NC-ND 3.0 US. (version: 29 June 2018).

Ayto, J. (2011). *Dictionary of Word Origins: The Histories of More Than 8,000 English-Language Words*. Arcade Publishing.

Bergsma, S. and Kondrak, G. (2007). Alignment-based discriminative string similarity. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 656–663.

Cresswell, J. (2010). *Oxford Dictionary of Word Origins*. Oxford University Press.

David Crystal, editor. (2011). *A Dictionary of Linguistics and Phonetics (6th ed.)*. David Blackwell Publishing. p. 104, ISBN 978-1-4443-5675-5. OCLC 899159900.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pages 67–71.

Jäger, G., List, J.-M., and Sofroniev, P. (2017). Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the*

European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1205–1216, Valencia, Spain, April. Association for Computational Linguistics.

Lewis, D. (2013). *Now I Know: The Revealing Stories Behind the World's Most Interesting Facts*. Adams Media.

List, J.-M., Walworth, M., Greenhill, S. J., Tresoldi, T., and Forkel, R. (2018). Sequence Comparison in Computational Historical Linguistics Phonetic Alignments and Cognate Detection with LingPy 2.6. *Journal of Language Evolution*.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2010). Quantitative analysis of culture using millions of digitized books. *Science*.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. In *Proceedings of the Royal Society of London*, pages 240–242.

Pechenick, E. A., Danforth, C. M., and Dodds, P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041.

Rabinovich, E., Tsvetkov, Y., and Wintner, S. (2018). Native language cognate effects on second language lexical choice. *CoRR*, abs/1805.09590.

Saussure, F. d. (1916). Cours de linguistique générale, ed. *C. Bally and A. Sechehaye, with the collaboration of A. Riedlinger, Lausanne and Paris: Payot*.

Tahmasebi, N., Borin, L., and Jatowt, A. (2018). Survey of computational approaches to lexical semantic change. *arXiv preprint arXiv:1811.06278*.

Uban, A., Ciobanu, A. M., and Dinu, L. P. (2019). Studying laws of semantic divergence across languages using cognate sets. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 161–166.